

Information extraction

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction.

Due to the difficulty of the problem, current approaches to IE focus on narrowly restricted domains. An example is the extraction from newswire reports of corporate mergers, such as denoted by the formal relation:

MergerBetween(*company*₁, *company*₂, *date*) ,

from an online news sentence such as:

"Yesterday, New York based Foo Inc. announced their acquisition of Bar Corp."

A broad goal of IE is to allow computation to be done on the previously unstructured data. A more specific goal is to allow logical reasoning to draw inferences based on the logical content of the input data. Structured data is semantically well-defined data from a chosen target domain, interpreted with respect to category and context.

Information Extraction is the part of a greater puzzle which deals with the problem of devising automatic methods for text management, beyond its transmission, storage and display. The discipline of information retrieval (IR)^[1] has developed automatic methods, typically of a statistical flavor, for indexing large document collections and classifying documents. Another complementary approach is that of natural language processing (NLP) which has solved the problem of modelling human language processing with considerable success when taking into account the magnitude of the task. In terms of both difficulty and emphasis, IE deals with tasks in between both IR and NLP. In terms of input, IE assumes the existence of **a set of documents in which each document follows a template, i.e. describes one or more entities or events in a manner that is similar to those in other documents but differing in the details**. An example, consider a group of newswire articles on Latin American terrorism with each article presumed to be based upon one or more terroristic acts. We also define for any given IE task a template, which is a(or a set of) case frame(s) to hold the information contained in a single document. For the terrorism example, a template would have slots corresponding to the perpetrator, victim, and weapon of the terroristic act, and the date on which the event happened. An IE system for this problem is required to “understand” an attack article only enough to find data corresponding to the slots in this template.

Contents

History

Present significance

Tasks and subtasks

World Wide Web applications

Approaches

Free or open source software and services

Commercial software and services

See also

References

External links

History

Information extraction dates back to the late 1970s in the early days of NLP.^[2] An early commercial system from the mid-1980s was JASPER built for Reuters by the Carnegie Group with the aim of providing real-time financial news to financial traders.^[3]

Beginning in 1987, IE was spurred by a series of Message Understanding Conferences. MUC is a competition-based conference^[4] that focused on the following domains:

- MUC-1 (1987), MUC-2 (1989): Naval operations messages.
- MUC-3 (1991), MUC-4 (1992): Terrorism in Latin American countries.
- MUC-5 (1993): Joint ventures and microelectronics domain.
- MUC-6 (1995): News articles on management changes.
- MUC-7 (1998): Satellite launch reports.

Considerable support came from the U.S. Defense Advanced Research Projects Agency (DARPA), who wished to automate mundane tasks performed by government analysts, such as scanning newspapers for possible links to terrorism.

Present significance

The present significance of IE pertains to the growing amount of information available in unstructured form. Tim Berners-Lee, inventor of the world wide web, refers to the existing Internet as the web of *documents* ^[5] and advocates that more of the content be made available as a web of data.^[6] Until this transpires, the web largely consists of unstructured documents lacking semantic metadata. Knowledge contained within these documents can be made more accessible for machine processing by means of transformation into relational form, or by marking-up with XML tags. An intelligent agent monitoring a news data feed requires IE to transform unstructured data into something that can be reasoned with. A typical application of IE is to scan a set of documents written in a natural language and populate a database with the information extracted.^[7]

Tasks and subtasks

Applying information extraction to text is linked to the problem of text simplification in order to create a structured view of the information present in free text. The overall goal being to create a more easily machine-readable text to process the sentences. Typical subtasks of IE include:

- **Named entity extraction** which could include:
 - **Named entity recognition**: recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions, by employing existing knowledge of the domain or information extracted from other sentences. Typically the recognition task involves assigning a unique identifier to the extracted entity. A simpler task is *named entity detection*, which aims at detecting entities without having any existing knowledge about the entity instances. For example, in processing the sentence "M. Smith likes fishing", *named entity detection* would denote **detecting** that the phrase "M. Smith" does refer to a person, but without necessarily having (or using) any knowledge about a certain *M. Smith* who is (or, "might be") the specific person whom that sentence is talking about.
 - **Coreference resolution**: detection of coreference and anaphoric links between text entities. In IE tasks, this is typically restricted to finding links between previously-extracted named entities. For example, "International Business Machines" and "IBM" refer to the same real-world entity. If we take the two sentences "M. Smith likes fishing. But he doesn't like biking", it would be beneficial to detect that "he" is referring to the previously detected person "M. Smith".
 - **Relationship extraction**: identification of relations between entities, such as:
 - PERSON works for ORGANIZATION (extracted from the sentence "Bill works for IBM.")
 - PERSON located in LOCATION (extracted from the sentence "Bill is in France.")
- **Semi-structured information extraction** which may refer to any IE that tries to restore some kind of information structure that has been lost through publication, such as:
 - **Table extraction**: finding and extracting tables from documents.
 - **Comments extraction** : extracting comments from actual content of article in order to restore the link between author of each sentence
- **Language and vocabulary analysis**
 - **Terminology extraction**: finding the relevant terms for a given corpus
- **Audio extraction**
 - **Template-based music extraction**: finding relevant characteristic in an audio signal taken from a given repertoire; for instance ^[8] time indexes of occurrences of percussive sounds can be extracted in order to represent the essential rhythmic component of a music piece.

Note that this list is not exhaustive and that the exact meaning of IE activities is not commonly accepted and that many approaches combine multiple sub-tasks of IE in order to achieve a wider goal. Machine learning, statistical analysis and/or natural language processing are often used in IE.

IE on non-text documents is becoming an increasingly interesting topic in research, and information extracted from multimedia documents can now be expressed in a high level structure as it is done on text. This naturally leads to the fusion of extracted information from multiple kinds of documents and sources.

World Wide Web applications

IE has been the focus of the MUC conferences. The proliferation of the Web, however, intensified the need for developing IE systems that help people to cope with the enormous amount of data that is available online. Systems that perform IE from online text should meet the requirements of low cost, flexibility in development and easy adaptation to new domains. MUC systems fail to meet those criteria. Moreover, linguistic analysis performed for unstructured text does not exploit the HTML/XML tags and the layout formats that are available in online texts. As a result, less linguistically intensive approaches have been developed for IE on the Web using wrappers, which are sets of highly accurate rules that extract a particular page's content. Manually developing wrappers has proved to be a time-consuming task, requiring a high level of expertise. Machine learning techniques, either supervised or unsupervised, have been used to induce such rules automatically.

Wrappers typically handle highly structured collections of web pages, such as product catalogs and telephone directories. They fail, however, when the text type is less structured, which is also common on the Web. Recent effort on *adaptive information extraction* motivates the development of IE systems that can handle different types of text, from well-structured to almost free text -where common wrappers fail- including mixed types. Such systems can exploit shallow natural language knowledge and thus can be also applied to less structured texts.

A recent development is Visual Information Extraction,^{[9][10]} that relies on rendering a webpage in a browser and creating rules based on the proximity of regions in the rendered web page. This helps in extracting entities from complex web pages that may exhibit a visual pattern, but lack a discernible pattern in the HTML source code.

Approaches

Three standard approaches are now widely accepted:

- Hand-written regular expressions (or nested group of regular expressions)
- Using classifiers
 - Generative: naïve Bayes classifier
 - Discriminative: maximum entropy models such as Multinomial logistic regression
- Sequence models
 - Hidden Markov model
 - Conditional Markov model (CMM) / Maximum-entropy Markov model (MEMM)
 - Conditional random fields (CRF) are commonly used in conjunction with IE for tasks as varied as extracting information from research papers^[11] to extracting navigation instructions.^[12]
- Human-in-the-loop
 - Text nailing

Numerous other approaches exist for IE including hybrid approaches that combine some of the standard approaches previously listed.

Free or open source software and services

- General Architecture for Text Engineering (GATE) is bundled with a free Information Extraction system

- [Apache OpenNLP](#) is a Java machine learning toolkit for natural language processing
- [OpenCalais](#) is an automated information extraction web service from [Thomson Reuters](#) (Free limited version)
- [Machine Learning for Language Toolkit \(Mallet\)](#) is a Java-based package for a variety of natural language processing tasks, including information extraction.
- [DBpedia Spotlight](#) is an open source tool in Java/Scala (and free web service) that can be used for named entity recognition and [name resolution](#).
- [Natural Language Toolkit](#) is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language
- [DeepDive \(http://deepdive.stanford.edu/\)](http://deepdive.stanford.edu/) is a system to extract value from dark data, created by a team from the InfoLab at [Stanford University](#).
- [geoparsepy \(https://pypi.python.org/pypi/geoparsepy\)](https://pypi.python.org/pypi/geoparsepy) is a free Python geoparsing library for location entity identification and disambiguation.
- [Read the Web \(http://rtw.ml.cmu.edu/rtw\)](http://rtw.ml.cmu.edu/rtw) is a machine learning system that acquires the ability to extract structured information from unstructured web pages.
- [OpenIE \(https://github.com/dair-iitd/OpenIE-standalone\)](https://github.com/dair-iitd/OpenIE-standalone) is a program that "creates extractions that represent relations in text."
- [Stanford OpenIE \(https://nlp.stanford.edu/software/openie.html\)](https://nlp.stanford.edu/software/openie.html) is a java program for the extraction of relation tuples, typically binary relations, from plain text.
- [CogComp Relation Extraction \(https://github.com/CogComp/cogcomp-nlp/tree/master/relation-extraction\)](https://github.com/CogComp/cogcomp-nlp/tree/master/relation-extraction) application which currently only detects relation pairs within the same sentence.
- See also [CRF implementations](#)

Commercial software and services

- IBM Watson^{[13][14]}
- Wolfram Natural Language Understanding^{[13][15]}

See also





- [Ontology extraction](#)
- [Applications of artificial intelligence](#)
- [Concept mining](#)
- [DARPA TIPSTER Program](#)
- [Enterprise search](#)
- [Faceted search](#)
- [Knowledge extraction](#)
- [Named entity recognition](#)
- [Nutch](#)
- [Semantic translation](#)
- [Textmining](#)
- [Web scraping](#)
- [Open information extraction](#)

Lists

- [List of emerging technologies](#)

- [Outline of artificial intelligence](#)

References

1. FREITAG, DAYNE. "Machine Learning for Information Extraction in Informal Domains" (<http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Seminar/freitag2000-ml.pdf>) (PDF). 2000 Kluwer Academic Publishers. Printed in The Netherlands.
2. Andersen, Peggy M.; Hayes, Philip J.; Huettner, Alison K.; Schmandt, Linda M.; Nirenburg, Irene B.; Weinstein, Steven P. "Automatic Extraction of Facts from Press Releases to Generate News Stories". [CiteSeerX 10.1.1.14.7943](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.7943) (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.7943>).
3. Cowie, Jim; Wilks, Yorick. "Information Extraction". [CiteSeerX 10.1.1.61.6480](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.6480) (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.6480>).
4. Marco Costantino, Paolo Coletti, Information Extraction in Finance, Wit Press, 2008. ISBN 978-1-84564-146-7
5. "Linked Data - The Story So Far" (<http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>) (PDF).
6. "Tim Berners-Lee on the next Web" (http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html).
7. R. K. Srihari, W. Li, C. Niu and T. Cornell, "InfoXtract: A Customizable Intermediate Level Information Extraction Engine", [Journal of Natural Language Engineering](http://journals.cambridge.org/action/displayIssue?iid=359643) (<http://journals.cambridge.org/action/displayIssue?iid=359643>), Cambridge U. Press, 14(1), 2008, pp.33-69.
8. A.Zils, F.Pachet, O.Delerue and F. Gouyon, Automatic Extraction of Drum Tracks from Polyphonic Music Signals (<http://www.csl.sony.fr/downloads/papers/2002/ZilsMusic.pdf>), Proceedings of WedelMusic, Darmstadt, Germany, 2002.
9. Chenthamarakshan, Vijil; Desphande, Prasad M; Krishnapuram, Raghu; Varadarajan, Ramakrishnan; Stolze, Knut. "WYSIWYE: An Algebra for Expressing Spatial and Textual Rules for Information Extraction". [arXiv:1506.08454](http://arxiv.org/abs/1506.08454) (<http://arxiv.org/abs/1506.08454>).
10. Baumgartner, Robert; Flesca, Sergio; Gottlob, Georg. "Visual Web Information Extraction with Lixto". [CiteSeerX 10.1.1.21.8236](https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.8236) (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.8236>).
11. Peng, F.; McCallum, A. (2006). "Information extraction from research papers using conditional random fields☆". *Information Processing & Management*. 42 (4): 963. doi:10.1016/j.ipm.2005.09.002 (<https://doi.org/10.1016/j.ipm.2005.09.002>).
12. Shimizu, Nobuyuki; Hass, Andrew (2006). "Extracting Frame-based Knowledge Representation from Route Instructions" (<http://www.cs.albany.edu/~shimizu/shimizu+haas2006frame.pdf>) (PDF).
13. Jiang, Jing (2012). "Information Extraction from Text" (http://www.stat.osu.edu/~dmsl/Information_Extraction.pdf) (PDF). Ohio State University Department of Statistics. Retrieved July 13, 2016.
14. "IBM Watson Information" (http://researcher.watson.ibm.com/researcher/view_group.php?id=1264). IBM. Retrieved July 13, 2016.
15. "Wolfram Data Framework: Take Data and Make It Meaningful" (<http://www.wolfram.com/data-framework/>). *www.wolfram.com*. Retrieved 2016-07-13.

External links

- [MUC](http://www.itl.nist.gov/iaui/894.02/related_projects/muc/) (http://www.itl.nist.gov/iaui/894.02/related_projects/muc/)
- [ACE](http://projects ldc.upenn.edu/ace/) (<http://projects ldc.upenn.edu/ace/>) (LDC)
- [ACE](http://www.itl.nist.gov/iad/894.01/tests/ace/) (<http://www.itl.nist.gov/iad/894.01/tests/ace/>) (NIST)

- [Alias-I "competition" page \(http://alias-i.com/lingpipe/web/competition.html\)](http://alias-i.com/lingpipe/web/competition.html) A listing of academic toolkits and industrial toolkits for natural language information extraction.
 - [Gabor Melli's page on IE \(http://www.gabormelli.com/RKB/Information_Extraction_Task\)](http://www.gabormelli.com/RKB/Information_Extraction_Task) Detailed description of the information extraction task.
 - [CRF++: Yet Another CRF toolkit \(https://web.archive.org/web/20100421020327/http://crfpp.sourceforge.net/\)](https://web.archive.org/web/20100421020327/http://crfpp.sourceforge.net/)
 - [A Survey of Web Information Extraction Systems \(http://www.csie.ncu.edu.tw/~chia/pub/iesurvey2006.pdf\)](http://www.csie.ncu.edu.tw/~chia/pub/iesurvey2006.pdf) A comprehensive survey.
 - [A multilingual corpus of news annotated with event information \(http://nlp.kiv.zcu.cz/projects/mevex\)](http://nlp.kiv.zcu.cz/projects/mevex)
-

Retrieved from "https://en.wikipedia.org/w/index.php?title=Information_extraction&oldid=822439233"

This page was last edited on 26 January 2018, at 11:14.

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.